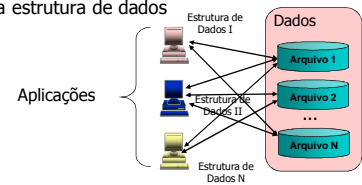
Mineração de Dados em Biologia Molecular
Bases de Dados
André C. P. L. F. de Carvalho
Monitor: Valéria Carvalho

Aula de hoje
Motivação
SGBD
Bases (bancos) de dados
Desenvolvimento de sistemas de bases de dados
Modelo relacional
Bases de dados biológicos
16/08/2012
André de Carvalho - ICMC/USP
2

Motivação
Computadores são usados para manipular e armazenar dados
Poucos dados: arquivos simples
Grandes volumes de dados?
16/08/2012
André de Carvalho - ICMC/USP
3

Motivação
Primeira opção: Sistemas de informação baseados em gerenciamento de arquivos
Rotinas específicas para tarefas específicas
Dados armazenados em disco, usando uma determinada estrutura de dados
Estrutura de Dados I
Dados
Arquivo 1
Aplicações
Estrutura de Dados II
Arquivo 2
...
Arquivo N
Estrutura de Dados N
16/08/2012
André de Carvalho - ICMC/USP
4

Motivação
Problemas:
Redundância e inconsistência de dados
Múltiplos formatos de arquivos
Duplicação de informações em arquivos diferentes
Dificuldade de acesso aos/manipulação dos dados
Novo programa precisa ser escrito para realizar cada nova tarefa
Problemas de integridade
Restrições de integridade ficam escondidas no código, ao invés de explicitamente indicadas
Ex.: saldo da conta > 0
16/08/2012
André de Carvalho - ICMC/USP
5

Motivação
Problemas
Falhas nas atualizações
Podem deixar a base de dados (BD) em um estado inconsistente, com atualizações apenas parciais
Ex.: transferência de recursos de uma conta para outras deve ou ser completa ou não ocorrer
Acesso concorrente por múltiplos usuários
Acesso concorrente sem controle pode levar a inconsistências
Dois usuários consultando e atualizando um arquivo ao mesmo tempo
Segurança
Difícil prover acesso a apenas parte dos dados
16/08/2012
André de Carvalho - ICMC/USP
6

Código	Sexo	Data de nascimento	País	Código da amostra	Região do corpo	Data de coleta	Código da seqüência	Região	Formato FASTA
1	M	1/1/1970	Brasil	1	fígado	10/10/2005	1111	ENV	'atgcctgacttgctacccttggaaatcga
1	M	1/1/1970	Brasil	1	fígado	10/10/2005	1112	POT	atgcctgacttgctacccttggaaacttaaa'
1	M	1/1/1970	Brasil	1	fígado	10/10/2005	1113	Null	aaatcga...tactttactttaccataacttaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2222	Null	accataacttaaa accataacttaaa
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	tacttactttactttttactttaccataacttaaa
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	tacttactttactttttactttaccataacttaaa
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3333	ENV	'atgcctgacta...tactttactttaccataacttaaa
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3334	Null	aaatcga...tactttactttaccataacttaaa'
2040	M	2/2/1980	Angola	5	pulmão	2/2/2006	3335	POT	aaatcga...tactttactttaccataacttaaa'

Código	Sexo	Data de nascimento	País	Código da amostra	Região do corpo	Data de coleta	Código da seqüência	Região	Formato FASTA
1	M	1/1/1970	Brasil	1	fígado	10/10/2005	1111	ENV	'atgcctgacttgctacccttggaaatcga
1	M	1/1/1970	Brasil	1	fígado	10/10/2005	1112	POT	atgcctgacttgctacccttggaaacttaaa'
1	M	1/1/1970	Brasil	1	fígado	10/10/2005	1113	Null	aaatcga...tactttactttaccataacttaaa'
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2222	Null	accataacttaaa accataacttaaa
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	tacttactttactttttactttaccataacttaaa
1	M	1/1/1970	Brasil	2	baço	5/5/2005	2223	POT	tacttactttactttttactttaccataacttaaa
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3333	ENV	'atgcctgacta...tactttactttaccataacttaaa
2040	F	2/2/1980	Angola	5	pulmão	2/2/2006	3334	Null	aaatcga...tactttactttaccataacttaaa'
2040	M	2/2/1980	Angola	5	pulmão	2/2/2006	3335	POT	aaatcga...tactttactttaccataacttaaa'

Arquivos tradicionais	SGBD
Definições dos dados é parte do código dos programas	Meta-dados
Dados e aplicação são dependentes	Dados e aplicação são independentes
Dados representados no nível físico	Representação conceitual
Cada módulo implementa uma visão dos dados	Múltiplas visões dos dados

Base de Dados

- Definição:
  - Coleção de dados logicamente relacionados que tem algum propósito associado
- Projetada, construída e preenchida com dados para satisfazer um propósito ou público específico
- Representa algum aspecto do mundo real
  - Mini-mundo

SGBD

- Tem funções para **definir (incluir), recuperar, excluir** e **modificar** dados em uma BD

SGBD

- Informação no sistema de BD deve ter uma estrutura
- Descrição da estrutura = esquema
  - Mantida na BD como metadados
- Conceitos e associações do mundo real devem ser capturados do mini-mundo e armazenados como metadados do SGBD
  - Exemplo:
    - Mini-mundo: universidade
    - Conceitos: cursos, disciplinas, aulas, alunos
    - Associações: cursos tem disciplinas, alunos se matriculam em cursos

Exemplo

codPaciente	sexo	nascimento	país
0001	M	1/1/1970	Brasil
1119	M	1/4/1970	Brasil
1209	F	16/10/1970	Chile
0002	M	21/1/1997	EUA
1987	F	15/1/1979	Brasil
1111	F	1/1/1980	Angola
2040	F	2/2/1980	Angola

SQL

- Structure Query language
- É uma linguagem de computador que segue o padrão ANSI
- SQL permite:
  - Consultar uma BD relacional
  - Recuperar dados de uma BD
  - Inserir novos registros em uma BD
  - Deletar registros de uma BD
  - Atualizar registros em uma BD

SGBDs que usam SQL

- Produtos comerciais
  - Microsoft ACCESS (Microsoft Office)
  - Microsoft SQLserver
  - Oracle
- Freeware
  - **MySQL**
  - PostgreSQL
  - MiniSQL

Bases de dados biológicos

- Gerais
  - Sequências de DNA, funções de proteinas, estruturas 3-dimensionais de proteínas, ...
- Especializados
  - EST, STS, SNP, RNA, genomas, famílias de proteina, pathways, dados de microarray, ...

BD de genomas

- Toda seqüência de genoma publicada deve ser disponibilizada em uma BD pública
  - Membros do International Nucleotide Sequence Database Collaboration são os principais repositórios
  - Consórcio formado por 3 grandes BDs
    - EMBL (European Molecular Biology Laboratory nucleotide sequence database at EBI, Hinxton, UK)
    - GenBank (at National Center for Biotechnology information, NCBI, Bethesda, MD, USA)
    - DDBJ (DNA Data Bank Japan at CIB , Mishima, Japan)

Conclusão

- Motivação
- SGBD
- Bases de dados
- Desenvolvimento de Sistemas de BDs
  - Modelo Relacional
- BD Biológicos

Agradecimento

- Prof Carlos Eduardo Ferreira, IME-USP

Perguntas